

NICAR Tip Sheet: How and why to make your data analysis reproducible

Short-link to this document: <http://bit.ly/nicar-2019-reproducibility>

Session link: <https://www.ire.org/events-and-training/event/3433/4167/>

Date/Time/Place

Thursday, March 7, 2019

9 a.m. to 10 a.m.

Cardiff

Panelists

Hannah Cushman

Ryann Grochowski Jones

Andrew Ba Tran

Jeremy Singer-Vine (moderator)

Panel Description

You understand how you processed your data. Does your editor? Your reader? You, in six months? Without a replicable approach to extracting, transforming and loading data, we are often frustrated in our efforts to share or update our work. Join us for a panel discussion of reproducible data workflows. We'll talk about why we use standardized processes for collecting, cleaning and analyzing data, and share practices that work for us. We'll also discuss strategies for smart human intervention (i.e. reporting, logging and documentation) in automated workflows.

Like what you heard? Highlighted examples from the session:

- “How We Analyzed Commercial and Industrial Property Assessments in Chicago and Cook County” (ProPublica Illinois): <https://projects.propublica.org/graphics/the-tax-divide-analysis> & <https://github.com/propublica/propertyassessments>
- “Making Data, the DataMade Way”: <https://github.com/datamade/data-making-guidelines>
- DataMade’s analysis of Chicago lead-in-water data for the City Bureau: <https://github.com/City-Bureau/chicago-lead>
- Andrew Tran’s r-journalism.com: <https://r-journalism.com/>

- The Washington Post's immigrant visas analysis:
 - https://github.com/wpinvestigative/immigrant_visas
 - https://www.washingtonpost.com/national/how-trump-is-changing-the-face-of-legal-immigration/2018/07/02/477c78b2-65da-11e8-99d2-0d678ec08c2f_story.html
 - <https://dailycaller.com/2018/07/16/center-immigration-studies-washington-post/>
 - <https://cis.org/Vaughan/Washington-Post-Analysis-Immigrant-Admissions-Under-Trump-Ignores-Half-All-Immigrants-Reach>
- The Washington Post's Global Terrorism Database analysis:
 - https://github.com/wpinvestigative/global_terrorism_database_analysis
 - https://www.washingtonpost.com/national/in-the-united-states-right-wing-violence-is-on-the-rise/2018/11/25/61f7f24a-deb4-11e8-85df-7a6b4d25cfbb_story.html
- "How We Approached Data Cleaning for Our Listeners' Favorite Albums of 2017" (NPR):
<http://blog.apps.npr.org/2018/01/03/all-songs-considered-poll.html>

More to chew on:

- HarvardX MOOC on reproducible data science:
<https://www.edx.org/course/principles-statistical-computational-harvardx-ph527x>
- Mike Bostock's "Why Use Make?" : <https://bost.ocks.org/mike/make/>
- "Muck: A Build Tool for Data Journalists" (Come for the promise of one build tool, stay for the software equivalent of a lit review of other build tools);
https://www.cjr.org/tow_center_reports/muck-a-build-tool-for-data-journalists.php
- Choosing a toolkit for shareable data analysis:
<https://source.opennews.org/articles/black-box-be-gone-tools-human-optimized-data-analy/> –
- A 50+ dimensional comparison of Luigi, Airflow, and Pinball:
<http://bytepawn.com/luigi-airflow-pinball.html>
- "How Open Should Open Source Data Visualization Be?" (Flowing Data):
<https://flowingdata.com/2008/08/29/a-case-for-open-source-data-visualization/>
- FiveThirtyEight's datasets: <https://data.fivethirtyeight.com/>
- An index of all BuzzFeed News open-source data, analysis, libraries, tools, and guides:
<https://github.com/BuzzFeedNews/everything>

- An Aug. 2015 Source article explaining BuzzFeed News' approach to reproducible data analysis; <https://source.opennews.org/articles/what-weve-learned-about-sharing-our-data-analysis/>

Thank you!